LÍNGUA NATURAL 2008/2009

Mini-Projecto Nº 2 — MP2

A realizar:	☐ individualmente	☒ **em grupo**
Local de trabalho:	☐ aula prática	☒ **casa (TPC)**
Local de entrega:	☐ aula teórica	☒ **submissão electrónica**
Data limite entrega:	**até às 12:00 (meio dia) do dia 27/Out**	

Objectivos operacionais

Construção de modelos de língua estatísticos, e sua utilização numa aplicação real.

Enunciado

1. Calcule os unigramas (ficheiro “unigramas0.txt”) e bigramas (ficheiro “bigramas0.txt”) referente às etiquetas presentes no ficheiro “text0.tagged”.

   Nota 1: Pode usar qualquer ferramenta para calcular os ficheiros de unigramas e bigramas.

   Nota 2: Para facilitar a tarefa de avaliação, os ficheiros calculados devem ter uma contagem por linha (ver os ficheiros “unigramas.txt” e “bigramas.txt”).

2. Calcule os unigramas e bigramas referentes às etiquetas presentes nos ficheiros “text1.tagged”, “text2.tagged”, “text3.tagged”, “text4.tagged”, “text5.tagged”, “text6.tagged” e “text7.tagged”;

3. Escreva um programa (P1) que compara a *proximidade* de dois ficheiros de N-gramas considerando que estes representam vectores: cada N-Grama corresponde a uma dimensão e o número de ocorrências à sua magnitude. Assim a proximidade entre dois vectores pode ser calculada usando a fórmula de semelhança do *cos*:

$$cos(d^a, d^b) = \frac{\sum_{i=1}^{n} d_i^a \times d_i^b}{\sqrt{\sum_{i=1}^{n} (d_i^a)^2} \times \sqrt{\sum_{i=1}^{n} (d_i^b)^2}}$$

   onde, *n* é a dimensão de cada vector *dn*.

   Nota: Não se esqueça que tem de normalizar as contagens, para que a dimensção dos textos não influencie os resultados obtidos.

4. Teste o programa P1 comparando os vectores definidos pelos unigramas referentes aos corpus de referência (“text0.txt”) e cada um dos outros textos (“text1.txt” a “text7.txt”);

5. Teste o programa P1 comparando os vectores definidos pelos bigramas referentes aos corpus de referência

("text0.txt) e cada um dos outros textos ("text1.txt" a "text7.txt");

6. Comente os resultados obtidos em 4 e 5.

7. Comente a viabilidade de desenvolver sistemas que indiquem se um documento está "bem escrito", com base em modelos de língua (N-gramas).

SUBMISSÃO

Envie, por e-mail, para o endereço oficial da disciplina (meic-ln@disciplinas.ist.utl.pt), um ficheiro zip (o nome do ficheiro deve ser formado por concatenação de "MP2-" com o número do grupo e com extensão ".zip") que deve conter:

- um ficheiro de texto (com o nome "opcoes.txt") com a descrição das opções tomadas, não podendo exceder 1 página A4;
- os ficheiro de texto com os unigramas ("unigramas0.txt" a "unigramas7.txt") e os ficheiros com os bigramas ("bigramas7.txt" a "bigramas7.txt") calculados a partir do corpus normalizado;
- o ficheiro de texto ("resultados.txt") com os resultados obtidos corresponte ao ponto 4 e 5 do enunciado;
- o ficheiro de texto ("comentarios.txt") com a análise corresponte ao ponto 6 do enunciado, não podendo exceder 1 página A4;
- o ficheiro de texto ("viabilidade.txt") com a análise corresponte ao ponto 7 do enunciado, não podendo exceder 1 página A4;
- um ficheiro de texto ("run.sh", ou "run.bat") com os comandos usados para obter todos os resultados reportados;
- Todo o código necessário à obtenção dos resultados apresentados.

Sempre que possível, todos os ficheiros devem conter a identificação do grupo e dos alunos participantes na elaboração deste trabalho.

CRITÉRIOS DE AVALIAÇÃO

Na avaliação serão tidos em conta os seguintes critérios:

1. Independência do sistema operativo;
2. Originalidade;
3. Cumprimento de todos os requisitos;
4. Correcção da soluções proposta;
5. Facilidade para proceder a alterações;
6. Cumprimento de todas as regras de submissão. O não cumprimento de qualquer regra implica um desconto mínimo de 2 valores.

CÓDIGO DE HONRA NA UNIVERSIDADE DE STANFORD (HTTP://WWW.STANFORD.EDU/DEPT/VPSA/JUDICIALAFFAIRS/GUIDING/HONORCODE.HTM)

The Honor Code is the University's statement on academic integrity written by students in 1921. It articulates University expectations of students and faculty in establishing and maintaining the highest standards in academic work:

1. The Honor Code is an undertaking of the students, individually and collectively:
    1. that they will not give or receive aid in examinations; that they will not give or receive unpermitted aid in class work, in the preparation of reports, or in any other work that is to be used by the instructor as the basis of grading;
    2. that they will do their share and take an active part in seeing to it that others as well as themselves uphold the spirit and letter of the Honor Code.
2. The faculty on its part manifests its confidence in the honor of its students by refraining from proctoring examinations and from taking unusual and unreasonable precautions to prevent the forms of dishonesty mentioned above. The faculty will also avoid, as far as practicable, academic procedures that create

temptations to violate the Honor Code.
3. While the faculty alone has the right and obligation to set academic requirements, the students and faculty will work together to establish optimal conditions for honorable academic work.

Examples of conduct that have been regarded as being in violation of the Honor Code include:

- Copying from another’s examination paper or allowing another to copy from one’s own paper
- Unpermitted collaboration
- Plagiarism
- Revising and resubmitting a quiz or exam for regrading, without the instructor’s knowledge and consent
- Giving or receiving unpermitted aid on a take-home examination
- Representing as one’s own work the work of another
- Giving or receiving aid on an academic assignment under circumstances in which a reasonable person should have known that such aid was not permitted

In recent years, most student disciplinary cases have involved Honor Code violations; of these, the most frequent arise when a student submits another’s work as his or her own, or gives or receives unpermitted aid. The standard penalty for a first offense includes a one-quarter suspension from the University and 40 hours of community service. In addition, most faculty members issue a "No Pass" or "No Credit" for the course in which the violation occurred. The standard penalty for multiple violations (e.g. cheating more than once in the same course) is a three-quarter suspension and 40 or more hours of community service.